FISHR Parameter Finder Documentation

Matthew C Keller
Associate Professor
Psychology and Neuroscience dept.
University of Colorado, Boulder

Doug Bjelland
Postdoctoral Trainee

Institute for Behavioral Genetics
1480 30th Street
Boulder, Colorado, 80303

August 25, 2016

1 Introduction

The `parameter_finder`2.4 utility program is provided with FISHR to help determine optimal input parameters for the `-ma_threshold` (which determines the IE moving average, MA, that optimizes IBD endpoints) and `-pie_threshold` (which determines the proportion of IEs [PIE], above which candidate IBD segments should be dropped). This utility program first requires the user to run GERMLINE2 (provided along with FISHR) using very conservative thresholds (e.g., a minimum cM threshold of 8 cM and 500 SNPs using only opposite homozygotes to determine endpoints); such a long cM threshold guarantees that the detected candidate segments (or at least the central most portion of them) are truly IBD. `parameter_finder`2.4 then uses as input the *bmatch, *bsid, *bmid files that are output by GERMLINE2, along with the original *ped file used as input to GERMLINE2. `parameter_finder`2.4

uses the middlemost portion of these IBD segments (specified using the `-cut-value` command) to infer the distribution of PIE and MA in truly IBD segments, and then compares these two distributions against those from random pairs of individuals matched at the same genomic locations. Users can then plot out and visually compare the PIE and MA distributions from truly IBD vs. random (ostensibly non-IBD) segments to find what they consider to be optimal parameters for their purposes. Users wanting fewer false negatives would choose thresholds that would catch almost all IBD segments at the expense of some rate of false calls, while those wanting fewer false positives would do the opposite.

Input Files

`-bmatch` [file path] output from GERMLINE2 that contains the information on the discovered IBD segments.

`-bsid` [file path] output from GERMLINE2 that contains information on the individuals in the dataset.

`-bmid` [file path] output from GERMLINE2 that contains the marker information.

`-ped-file` [file path] file containing the phased genotypes for all individuals in the dataset.

Parameters

`-window` [integer] the length, in SNPs, of the window used to calculate the moving average of implied errors (default is 50).

`-cut-value` [numeric] the proportion of the center section of the long IBD segments to use in determining optimal parameter. E.g., .60 would assume the middlemost 60% of each candidate segment is truly IBD, thereby only reporting MA and PIE statistics from the middlemost 60% of the original GERMLINE2 candidate segment (removing the first and last 20% of the original segments).

`-reduced` [integer][numeric] two values, the minimum length of the original candidate IBD segments in SNPs and the minimum length of the original candidate IBD segments in cM from GERMLINE2 (before any trimming from `-cut-value`), to use to determine the MA and PIE statistics from ostensibly truly IBD segments.

`-output-type` [character] the type of output for the program (Error1, Final, etc.; see FISHR manual). For almost all purposes of using this utility, users should use Error1 here.

`-log-file` [file path] the name of the log file to be printed out.

Output Files

Error1 columns: Individual 1, Individual 2, Start position of segment, End position of segment, length of segment (SNPs), length of segment (cM), Index of start SNP, Index of end SNP, PIE, Maximum moving average of implied errors (MA), PIE of random pair of individuals, Maximum MA of random pair of individuals, location of IEs within segment (separated by backslash).

Example

First, run GERMLINE2 for finding very long candidate IBD segments (Note: this can be done genome-wide, but for most purposes, just doing so on a subset of chromosomes or a subset of the total dataset is sufficient):

```
GERMLINE2 -pedfile Test.8k.ped -mapfile Test.8k.map -outfile Test.8k
-bin_out -err_hom 0 -err_het 0 -reduced -bits 50 -min_m 6 -w_extend
```

Explanation of GERMLINE2 input commands:
`-pedfile` - a PHASED pedfile.
`-mapfile` - a map file with cM distances in 3rd column
`-bin_out` - a compressed output necessary for being read by `parameter_finder` and FISHR

-err_hom 0 - allow 0 mismatching homozygous markers
-err_het 0 - allow 0 mismatching heterozygous markers
-reduced - a flag telling GERMLINE2 to reduce the columns of the output; necessary for being read by parameter_finder and FISHR
-bits 50 - the number of SNPs in each fixed window that GERMLINE2 uses for initial matches
-min_m - minimum length in cM for match to be output; here 6
-w_extend - extend match beyond the "bits" window until the first opposite homozygote (OH) occurs.
-h_extend - tells GERMLINE2 to use phase information as well as just opposite homozygosity in determining the SH endpoints. It is important NOT to use this option when running GERMLINE2 to find optimal parameters. This is because using phase information will bias the detected SHs to have fewer phase errors than randomly chosen truly IBD segments, and will make our sensitivity and specificity values based on the chosen thresholds appear better than they really would be in real circumstances. Although it's true that OHs also cause IEs, they make up a minority of IEs, most of which are caused by phase and SNP errors.

After running GERMLINE2 above, the user then runs parameter_finder2.4:

```
parameter_finder2.4 -bmatch Test.8k.bmatch -bsid Test.8k.bsid -bmid
Test.8k.bmid -ped-file Test.8k.ped -window 50 -cut-value .60 -reduced
300 8 -output-type Error1 -log-file pf_test.log | gzip -c > pf_test_out.gz
```

Explanation of parameter finder input commands:
-bmatch - the binary output (from text -bin_out) of GERMLINE2; has one SH per row
-bsid - the subject IDs output from GERMLINE2
-bmid - the marker IDs output from GERMLINE2
-ped-file - the path to the phased data; should be identical to the input -pedfile used in GERMLINE2
-window 50 - says to use a 50 SNP window for figuring moving average of IEs (MA)
-cut-value .6 - says to trim the SH to the middlemost 60%: 20% from the left and 20% from the right are trimmed. This ensure that the remaining segment is almost certainly IBD

`-reduced 500 8` - recommended parameters in real data; only use initial SHs that are at least 500 SNPs long and $>$ 8cM. The outputted middlemost segments will typically be $<$ 8cM (e.g., $\sim$ 4+ cM)
`-output.type Error1` - this is the typical output format to be requested.
`| gzip >` this pipes the standard out to gzip so the final output is a gzipped file.

FIGURES:

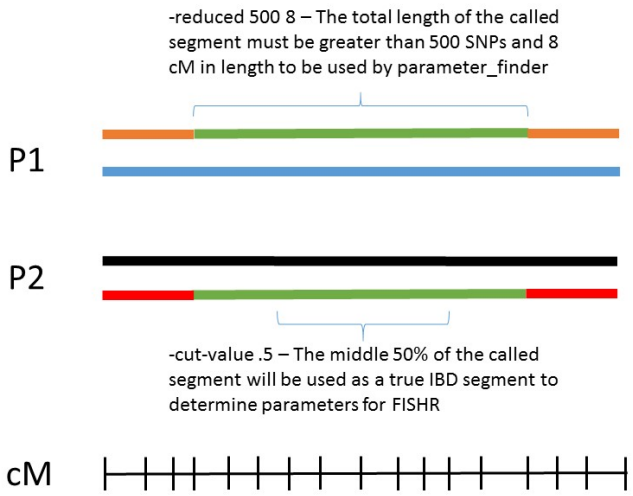

Figure 1: A called segment is displayed below showing the values for the $-reduced$ parameter and the $-cut-cm-value$ parameter. The first command allows for only the selection of very long called segments that we can be the most certain that they are IBD. The $-cut-value$ will then select the middle X% of the called segment. As we are uncertain of the exact end positions of the called segments at this point, we only use the very middle sections of the long segments to further increase the likelihood of what we are using is truly IBD.

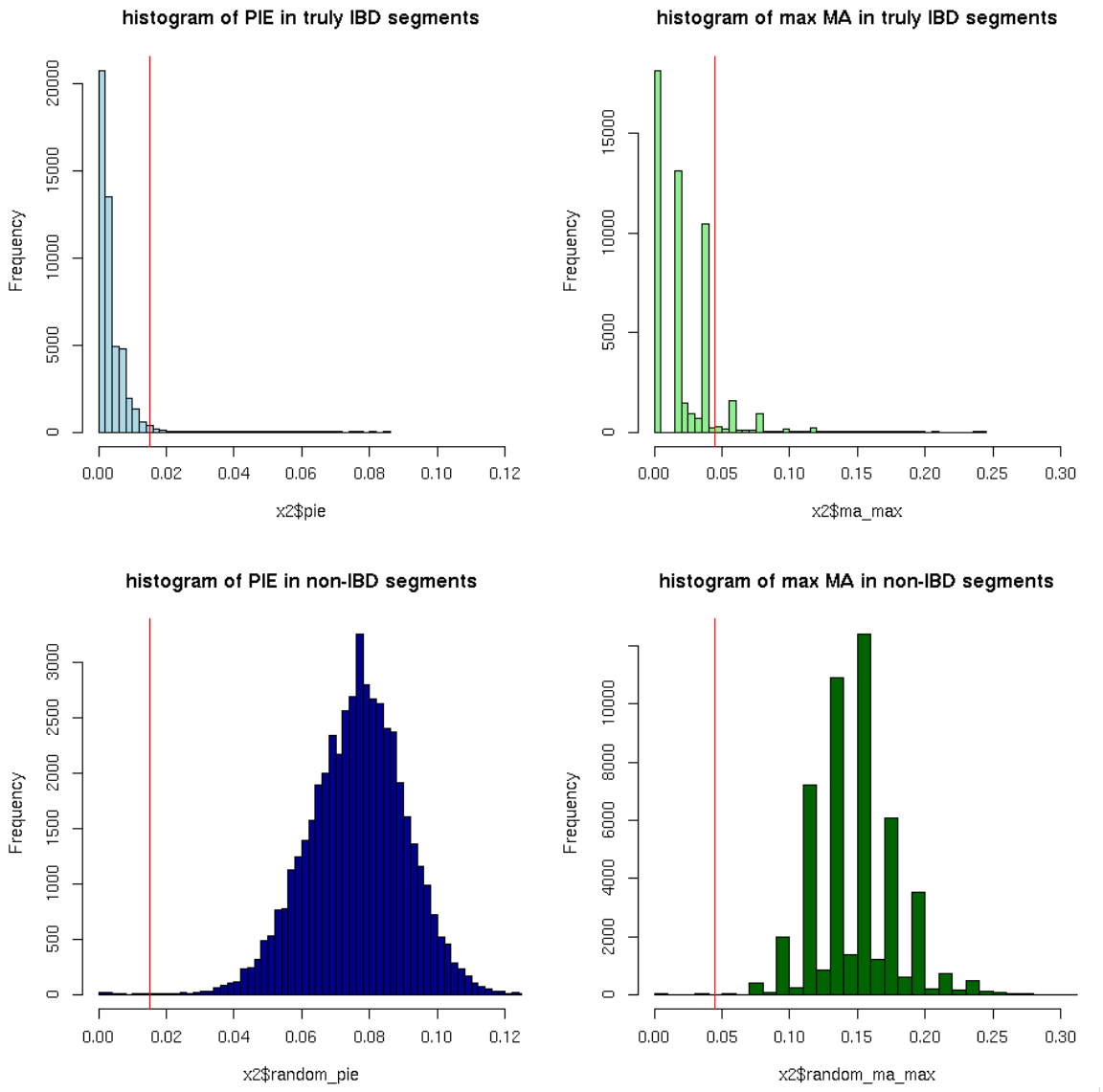

Figure 2: Comparison of PIE (left) and maximum MA (right) from truly IBD segments (top row) vs. non-IBD segments (bottom row). The red vertical lines show possible thresholds that might be chosen to lead to relatively high sensitivity and specificity.